Extraction de relations et de patrons de relations entre entités nommées en vue de l'enrichissement d'une ontologie

Yayoi Nakamura-Delloye, Rosa Stern

Extraction de relations et de patrons de relations entre entités nommées en vue de l'enrichissement d'une ontologie

Yayoi Nakamura-Delloye*, Rosa Stern* **

*ALPAGE, INRIA-Rocquencourt & Université Paris 7 Denis Diderot
Domaine de Voluceau Rocquencourt - B.P.105 78153 Le Chesnay
**Agence France Presse / Medialab, 13, place de la Bourse, 75002 Paris
yayoi@yayoi.fr
http://www.yayoi.fr
rosa.stern@afp.com

Résumé. Nous proposons dans cet article une méthode non-supervisée d'extraction des relations et des patrons de relations entre entités nommées, réalisée dans le cadre de la création et l'enrichissement d'une ontologie. La méthode proposée se caractérise par l'exploitation des résultats d'analyse syntaxique, notamment les chemins syntaxiques reliant deux entités nommées dans les arbres de dépendance. Les informations sur les relations syntaxiques présentes entre les composants sont mises à profit pour le calcul de la similarité employée pour la phase principale de classification. Nous présentons également le mécanisme conçu pour l'intégration des résultats obtenus dans une ontologie.

1. Introduction

L'organisation de connaissances à l'aide de ressources ontologiques constitue un enjeu important suscitant un intérêt croissant dans les travaux de recherche en traitement automatique des langues. Il peut s'agir de connaissances sur le monde pouvant aider des techniques de traitement automatique du langage (ci-après TAL) dans la tentative de compréhension automatique de textes ou d'applications exploitant les résultats de travaux en TAL dans le but d'enrichir de telles ressources. Le cas qui nous intéresse ici est celui d'un système permettant l'extraction de connaissances à partir de textes à l'aide du TAL, dans le but d'enrichir des ressources ontologiques préalablement constituées.

Il s'agit d'identifier dans de larges corpus textuels les relations pouvant exister entre des entités telles que des personnes et des organisations, appelées entités nommées (EN ci-après). On s'intéresse notamment aux relations d'appartenance entre ces deux types d'entités, ainsi qu'à d'autres qui seront présentées par la suite. Cette identification se fait à partir de résultats d'analyse syntaxique en dépendance des textes considérés. Cette identification peut se faire dans la perspective d'enrichir des connaissances sur un ensemble d'entités, organisées dans un référentiel ontologique.

Nous avons proposé dans (Nakamura-Delloye et al. 2010) une méthode d'acquisition semi-supervisée de relations entre entités nommées, basée sur un principe d'induction : quelques exemples de couples d'EN en une relation donnée, proposés par examen du corpus et introspection, sont fournis au système et permettent d'en extraire de nouvelles. L'inconvénient de cette méthode est la difficulté de déterminer les relations intéressantes qu'il est possible d'extraire à partir d'un corpus donné et de trouver des exemples pertinents de ces relations pour l'opération d'extraction. Ainsi, nous avons cette fois développé une méthode non-supervisée d'extraction et nous présentons dans cet article une évaluation de cette méthode.

Après avoir évoqué les recherches en relation avec notre problématique (§ 2), nous décrirons le cadre dans lequel nous proposons de mettre en œuvre nos travaux (§ 3). Puis nous détaillerons la procédure que nous proposons d'appliquer, des résultats de l'analyse syntaxique à l'extraction des relations et de leurs patrons (§ 4). La section suivante (§ 5) rend compte de l'expérience menée à partir de cette procédure et de son évaluation. Enfin, des perspectives d'intégration de ces travaux dans une ontologie seront également présentées (§ 6).

2. État de l'art

L'extraction de relations entre entités nommées est une opération importante pour beaucoup d'applications et de nombreuses études ont été proposées dans différents cadres de travail tels que la conception d'un système de question-réponse (Iftene et al. 2008), l'extraction d'information (Banko et al. 2007) ou l'extraction de réseaux sociaux (Matsuo 2006).

De nombreuses méthodes supervisées d'acquisition de relations basées sur de larges corpus annotés telles que (Zelenko et al. 2002), ont été proposées. L'utilisation de données annotées présente cependant le défaut majeur du coût très élevé de l'annotation manuelle. Des approches semi-supervisées ont donc été proposées, se fondant généralement sur un principe d'*induction* : on recourt à un ensemble réduit d'exemples du cas recherché. Cette approche initialement utilisée dans les travaux d'identification des patrons textuels, tels que ceux de Hearst (1992), a aussi été utilisée dans des travaux sur l'extraction des patrons de relations des EN (Brin 1998, Agichtein et al. 2000). Nous avons également proposé une méthode semi-supervisée (Nakamura-Delloye et al. 2010), basée sur ce principe d'induction. Mais ces méthodes semi-supervisées présentent à leur tour des inconvénients. Le manque de richesse dans le type de relation identifiée en est un : elles se limitent aux mêmes types que celles données en exemple. Mais leur défaut capital est la difficulté de déterminer des relations intéressantes pouvant être extraites à partir d'un corpus donné, et la difficulté de trouver des exemples de ces relations pertinentes pour l'opération d'extraction.

Hasegawa et al. (2004) propose une méthode non-supervisée évitant cet écueil, en se basant sur l'hypothèse que les couples d'EN en même relation apparaissent dans les mêmes contextes et que les mots représentatifs de leurs contextes peuvent caractériser leurs relations. Deux grandes étapes sont suivies : clustering des contextes puis étiquetage des clusters par extraction des mots représentatifs à partir de contextes. Différentes améliorations de cette approche ont été proposées par la suite (Zhang et al. 2005, Chen et al. 2005, He et al. 2006, Bollegala et al. 2010). La méthode que nous proposons dans cet article se fonde également sur ce principe, mais se distingue de travaux antérieurs notamment par l'exploitation de résultats d'analyse syntaxique et par son application précise à la constitution semi-automatique d'une ontologie qui prévoit une validation manuelle des données par des experts.

3. Procédure d'enrichissement de l'ontologie

Nous travaillons sur un corpus de dépêches de l'Agence France Presse, analysé syntaxiquement avec l'analyseur FRMG[1] (cf. 1 de la figure 1), qui emploie lui-même un module de reconnaissance d'entités nommées fournissant un étiquetage et un typage des entités[2].

L'identification des relations entre entités correspond à deux étapes principales : extraction et regroupement des couples d'EN et des chemins syntaxiques qui les relient (cf. 2 de la figure et § 4.1), puis acquisition des relations et des patrons correspondants (cf. 3 de la figure). Deux méthodes ont été expérimentées pour l'acquisition des relations et de leurs patrons : méthode semi-supervisée par induction (Nakamura-Delloye et al. 2010) et méthode non supervisée par clustering (§ 4.2). Suite à cette opération, on obtient des propositions de relations (cf. 4 de la figure) et des propositions de patrons (cf. 5 de la figure). Ces propositions sont fournies aux experts sous forme de « tickets » en vue de l'enregistrement dans la base après validation manuelle (cf. 6 de la figure et § 6).

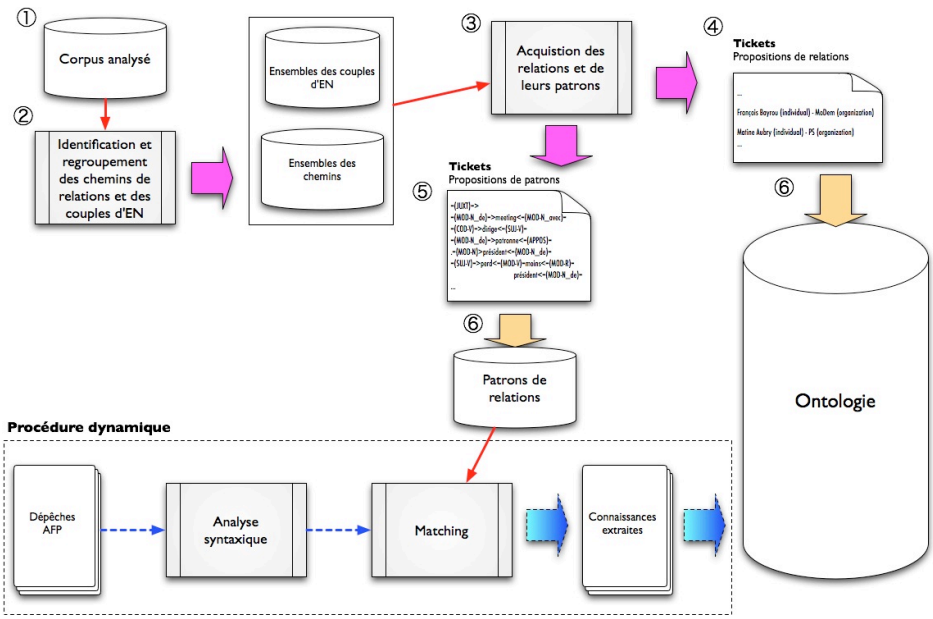

FIG. 1 – *Procédure générale d'extraction et d'enrichissement d'une ontologie*

[1] Pour la description de l'ensemble des analyseurs utilisés et leurs notations, voir notamment (De la Clergerie 2010) et (De la Clergerie et al. 2009).

[2] La détection des entités nommées est réalisée par SxPipe (Sagot & Boullier 2008, Stern & Sagot 2010-1 et 2010-2).

4. Extraction des relations et de leurs patrons

4.1 Première étape d'identification et de regroupement

La première étape de l'extraction (cf. 2 de la figure 1) consiste (1) en identification des couples d'EN et des chemins les reliant, et (2) en regroupement des couples d'EN reliées par les mêmes chemins syntaxiques.

Étant donné que les données initiales contiennent beaucoup d'informations de diverses natures, on extrait d'abord les seules données nécessaires à la construction de l'arbre syntaxique et celles sur les constituants de la phrase correspondant aux nœuds de l'arbre. Avec les données extraites du résultat d'analyse syntaxique, l'arbre syntaxique d'entrée est construit à partir de l'ensemble des relations de dépendance entre les constituants (cf. Figure 2).

4.1.1 Identification des couples d'EN et des chemins de relations

Notre première hypothèse a été, comme dans (Lin et al. 2001, Bunescu et al. 2005), que la relation entre deux EN était représentée dans l'arbre syntaxique par le chemin reliant les deux nœuds leur correspondant. Ainsi, dans l'arbre de la figure 2, la relation entre les deux EN, Eric Roy et Roger Ricort, est représentée par le chemin reliant leurs nœuds tracé par la ligne non contiguë, constitué d'une suite d'arcs et de nœuds intermédiaires : $\rightarrow_{\text{Sujet-V}}$ remplacera $\leftarrow_{\text{COD-V}}$. Nous appelons ces chemins « chemins syntaxiques de relations ».

Eric Roy remplacera officiellement au poste de directeur sportif Roger Ricort

```
W  1   Eric Roy      _PERSON_m
W  2   remplacera v
W  3   officiellement   adv
W  4   à    prep
W  5   le   det
W  6   poste   nc
W  7   de   prep
W  8   directeur   nc
W  9   sportif   adj
W  10  Roger Ricort   _PERSON3
R  SUJ-V   1   2
R  COD-V   10  2
R  CPL-V_à   6   2
R  MOD-V   3   2
R  MOD-N   9   8
R  MOD-N_de  8   6
```

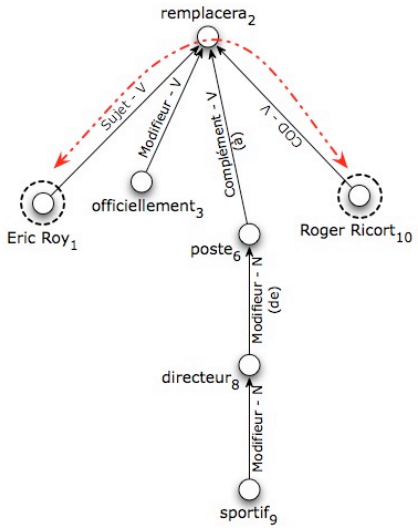

FIG. 2 – *Arbre syntaxique et chemin syntaxique de relation*

4.1.2 Regroupement des couples d'EN et des chemins de relations

Après avoir identifié et extrait tous les couples d'EN et les chemins les reliant dans les arbres syntaxiques, nous regroupons les couples d'EN qui partagent le même chemin de relation. L'exemple suivant est l'ensemble constitué des couples des EN_1, EN_2 partageant le même chemin de relation $\rightarrow_{MOD\text{-}N}$ qui signifie que EN_1 dépend syntaxiquement de EN_2 et qu'elles sont en relation Modifieur-Nom.

> Ensemble 1 : $\rightarrow_{MOD\text{-}N}$

Hasina (individual) - Ligue Awami (organization)

Jérôme Alonzo (individual) - PSG (organization)

De Gaulle (individual) - ORTF (organization)

Ibarretxe (individual) - Parti nationaliste basque (organization)

Metz (individual) - Parti communiste (organization)

...

4.2 Méthode non-supervisée d'acquisition des relations et des patrons de relations des entités nommées

À partir des ensembles des couples d'EN ainsi constitués, on extrait des relations intéressantes entre EN et leurs patrons. Nous proposons ici une méthode non-supervisée. L'acquisition se déroule en trois étapes : calcul de la similarité des chemins, classification des chemins et étiquetage des classes de chemins ainsi constituées.

4.2.1 Calcul de la similarité

La première étape d'acquisition consiste à calculer la similarité des chemins. À cet effet, les chemins sont représentés dans un espace vectoriel par leurs composants lexicaux (correspondant aux nœuds des arbres) pondérés avec la mesure tf.idf. Par ailleurs, nous avons apporté une amélioration à cette mesure pour notre tâche, par la prise en compte des relations syntaxiques existant entre les composants lexicaux. Les termes régissant les autres termes en relation avec eux sont considérés comme constituant le noyau sémantique et favorisés par rapport aux éléments dépendants. Ainsi, dans le chemin → filiale → rachat ←, le terme rachat qui régit deux éléments reçoit une pondération plus importante que filiale qui dépend de lui. Cette formule favorise, lors de la classification, le rapprochement de ce chemin avec la classe rachat plutôt qu'avec la classe filiale. En d'autres termes, la relation Fiat-General Motors reliée par ce chemin (*rachat par Fiat de la filiale allemande de General Motors*) rentre dans la classe rachat et non dans filiale. La valeur du mot i du chemin j est donc calculée comme suit :

$$m_i^j = tf_i^j \cdot idf_i \cdot p_i^j$$

tf_i^j correspond à la fréquence du mot dans le chemin, et idf_i est calculé à partir du nombre de chemins où le terme i n'apparaît pas, compte tenu du nombre total de chemins. Dans nos travaux, p est défini à 1 pour les termes régissants, à 0,8 sinon.

Les similarités entre les vecteurs α, β représentant les chemins sont ensuite calculées selon la similarité cosinus :

$$\cos\theta = \frac{\alpha \cdot \beta}{|\alpha| \cdot |\beta|}$$

4.2.2 Classification des chemins et étiquetage des classes

Dans la deuxième étape, le clustering des chemins est réalisé par une méthode de classification hiérarchique. Dans nos travaux, le saut minimum est adopté pour la mesure de dissimilarité inter-classe.

Les classes ainsi construites sont ensuite étiquetées par le terme apparaissant le plus fréquemment sur les chemins partagés comme dans (Hasegawa et al. 2004).

5. Expérience

Une expérience a été réalisée avec un corpus constitué d'un an de dépêches AFP (de janvier à décembre 2009), pour évaluer notre méthode d'acquisition non-supervisée. Après le pré-traitement qui élimine les phrases ne contenant pas plus d'une entité, le corpus contient 1 174 600 phrases et occupe 1,4Go de mémoire.

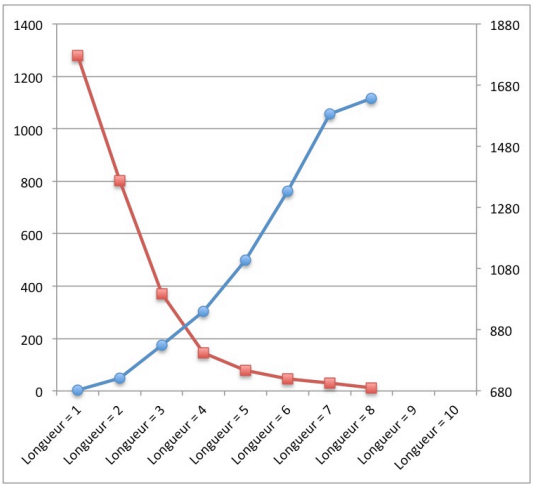

FIG. 3 – Productivité des chemins (points carrés) et temps de calcul (points ronds)

La longueur maximum d'un chemin est fixée à cinq et les couples d'EN reliées par un chemin plus long ne sont donc pas considérés. Cette valeur a été définie suite à l'étude de la productivité des chemins (i.e. nombre de couples d'EN qu'ils relient) selon la longueur et du temps de calcul selon la longueur maximum définie (cf. Fig. 3) : elle est jugée comme le seuil limite pour obtenir un nombre significatif de couples d'EN sans doubler le temps de calcul. De plus, nous n'avons traité que les chemins reliant au moins deux couples différents d'EN. En revanche, contrairement aux travaux antérieurs qui ne prenaient en compte que des couples ayant plus de trente cooccurrences, nous n'avons défini aucun seuil pour les couples, en supposant que l'utilisation des chemins syntaxiques permettait de repérer de manière efficace et fiable les couples d'EN en une certaine relation sémantique.

L'extraction des relations a été réalisée pour trois types de relations : relations Individu-Organisation (IND-ORG), Individu-Individu (IND-IND) et Compagnie-Compagnie (COM-COM).

5.1 Données obtenues

Le tableau suivant (TAB. 1) récapitule le volume de données que nous avons obtenues avec notre méthode d'extraction à partir du corpus décrit précédemment.

	IND-ORG	IND-IND	COM-COM
Couples EN	11 203	17 190	867
Couples EN (≥30)	370 (177)	73	21 (65)
Chemins	2476	3484	212
Classes	362 (34)	374	28 (15)
Couples classés	9380	9125	258
Chemins classés	1923	2112	76

TAB. 1 – *Nombres de classes, relations et chemins obtenus : les valeurs entre parenthèses sont les résultats de l'extraction présentés dans les travaux antérieurs (Hasegawa et al. 2004)*

La ligne « Couples EN » indique le nombre de couples d'EN extraits et la ligne « Couples EN (>30) », le nombre de couples de fréquence supérieure à 30. La ligne « Chemins » correspond au nombre de chemins extraits reliant au moins deux couples différents. Comme nous pouvons le constater, les couples COM-COM sont extrêmement restreintes dans notre résultat. Cela est probablement dû à la différence de nature de corpus, mais également à la performance de l'étiqueteur des entités nommées employé. Ce repérage limité des EN du type Compagnie a une influence cruciale sur le résultat d'extraction comme nous allons bientôt le décrire. Par

ailleurs, l'absence de seuil pour les couples d'EN nous a permis d'extraire un nombre beaucoup plus important de relations. Dans la procédure d'enrichissement de l'ontologie telle que nous l'envisageons, il est prévu qu'une phase de validation manuelle soit conduite par des experts. Il paraît ainsi préférable de privilégier le rappel à la précision, c'est-à-dire en augmentant le nombre de relations candidates. Notre intérêt réside donc dans l'augmentation du nombre de propositions avec une minimisation de la baisse de la précision.

La ligne « Classes » correspond au nombre de classes formées suite à la classification des chemins, comprenant plus d'un chemin. La ligne « Chemins classés » indique le nombre de chemins appartenant à une de ces classes et la ligne « Couples classés » correspond au nombre total de couples reliés par ces chemins classés et appartenant par conséquent à une classe. Dans la mesure où un couple peut être relié par différents chemins, il peut appartenir à plusieurs classes différentes. Ainsi, on peut, par exemple, trouver le couple (Xavier Bertrand, UMP) aussi bien dans la classe *porte-parole* que celle de *secrétaire (général)* ou encore *arrivée (à la tête)*. Dans les travaux de Hasegawa, un couple ne peut appartenir qu'à une seule classe, celle représentant la relation la plus représentative (*major relation*) pour le couple en question. Le choix entre ces possibilités pourrait être un sujet à débattre, mais il dépend sans doute aussi de l'application visée.

5.2 Méthode d'évaluation

Afin de nous munir de critères pour analyser quantitativement nos résultats, nous avons employé pour notre évaluation la méthode proposée dans les travaux antérieurs avec certaines adaptations. Il nous semble ici important de signaler que la comparaison de nos résultats nécessite certaines interprétations particulières, et la prise en compte du fait que nous avons une approche différente de celle des travaux antérieurs, puisque sans parler de l'utilisation ou non d'une analyse syntaxique, le corpus et les outils de prétraitement diffèrent.

5.2.1 Constitution des couples de référence

Pour l'évaluation, une étape de préparation s'impose : nous collectons les couples d'EN comptant plus de trente cooccurrences dans le corpus. Ces couples ont été ensuite vérifiés manuellement pour savoir si les EN formant les couples entretenaient effectivement une certaine relation jugée intéressante et pour déterminer, le cas échéant, la nature de leur relation. Le résultat de cette vérification manuelle donne lieu à une liste des couples « de référence ». Dans cette évaluation, nous avons constitué trois listes IND-ORG, IND-IND et COM-COM contenant respectivement 370, 73 et 65 couples de référence.

5.2.2 Mesures d'évaluation

Les rappel, précision et F-mesure sont également calculés comme suit :

Rappel : nombre de couples détectés parmi ceux qui figurent dans la liste des couples de référence, calculé par la formule suivante :

$$R = N_{correct} / N_{couples\_de\_référence}$$

Précision : nombre de couples corrects parmi l'ensemble des couples détectés, calculé par la formule suivante :

$$P = N_{correct} / (N_{correct} + N_{incorrect})$$

Pour compter les nombres de couples d'EN corrects et incorrects, nous avons choisi de manière aléatoire un nombre donné de chemins (200 chemins pour IND-ORG et IND-IND, et la totalité pour COM-COM) dont nous avons ensuite vérifié manuellement la validité selon deux aspects : (1) analyse syntaxique et (2) correspondance sémantique par rapport à la relation exprimée par leur classe. Le nombre de couples corrects correspond au nombre total des couples d'EN reliées par des chemins jugés corrects par la vérification manuelle. Le nombre de couples incorrects correspond quant à lui au nombre de ceux reliés par des chemins jugés incorrects.

F-mesure : score calculé par la combinaison du rappel et de la précision comme suit :

$$F = 2RP / (R + P)$$

5.3 Évaluation

	IND-ORG	IND-IND	COM-COM
Rappel	80 (83)	73	42 (74)
Précision	88 (79)	70	46 (76)
F-mesure	88 (80)	71	44 (75)

TAB. 2 – *Résultat d'évaluation : les valeurs entre parenthèses sont les résultats de l'extraction présentés dans les travaux antérieurs (Hasegawa et al. 2004)*

5.3.1 Description et remarques générales

Contrairement à ce à quoi nous nous étions attendus, le taux de rappel est assez bas. Cela va également à l'opposé de l'impression que nous avions eue lorsque nous avons vu le résultat avec un nombre important de propositions de relations. En effet, beaucoup de couples de fréquence supérieure à 30 sont reliés par des chemins sans

nœud intermédiaire, représentant la relation syntaxique réalisée par une apposition ou une juxtaposition. Notre méthode non-supervisée, basée notamment sur la classification des chemins par calcul de leur similarité lexicale, ignore complètement ces chemins. Cela a empêché le traitement des relations fréquentes réalisées principalement par ces chemins, ce qui a entraîné le taux de rappel obtenu. Mais, les relations sémantiques représentées par ces chemins telles que « appartenance » sont tellement larges que la prise en compte de ces chemins entraînerait la constitution d'une très grande classe englobant des sous-classes telles que « *est président de* », « *est porte-parole de* », qui sont utiles pour le peuplement de notre référentiel ontologique.

Pour les relations IND-ORG, nous avons vérifié 200 chemins, dont 132 étaient corrects. Mais la plupart des chemins erronés fournissaient peu de relations, ce qui a donné une précision élevée en termes de nombre de relations. Les relations IND-IND, beaucoup moins nombreuses (seulement un cinquième en termes de nombre de relations supérieures à 30 occurrences), ont cependant donné un résultat plus intéressant que ce que nous avions espéré. La non validité des chemins est souvent due aux erreurs d'analyse syntaxique et, dans la plupart des cas, les classes représentées par un terme qui n'implique pas intuitivement le type de relation traitée, contiennent des chemins créés par une fausse analyse syntaxique. Ainsi, les classes *député, ministre, chef* sont des classes comprenant des chemins non valides pour la relation IND-IND, alors qu'elles sont très productives et pertinentes pour la relation IND-ORG. Il y a aussi des classes constituées du fait d'erreurs d'étiquetage des EN. La formation des classes *directeur, capitaine, patron* dans le résultat de l'extraction des relations IND-IND est due au faux étiquetage des EN Organisation en EN Individu. Tous les chemins et les relations de ces classes sont valides pour les relations IND-ORG. Toutefois, nous n'avons pas pris en compte dans notre évaluation les erreurs dues aux fausses étiquettes d'EN, contrairement à celles provenant d'une fausse analyse syntaxique.

Avec un nombre restreint de données, l'extraction des relations COM-COM n'a pas donné de résultat satisfaisant. Afin de résoudre ce problème, nous avons essayé un algorithme qui permettait de prendre en compte également les chemins reliant seulement un couple d'EN, à condition tout de même que ce dernier soit relié au moins par un autre chemin. Cet algorithme a nécessité une validation manuelle supplémentaire – quoique très simple (une quinzaine de minutes) –, mais a donné un résultat intéressant, à savoir : rappel à 88, précision à 78 et F-mesure à 82. Cet algorithme pourrait proposer une solution alternative lorsque le volume de données d'entrée est limité.

Dans nos résultats, le nombre de classes formées est relativement important, mais avec une modification de la formule de calcul de similarité des chemins pour la classification, nous pouvons envisager la réduction de ce nombre de classes formées par augmentation des agrégations lors de la classification.

5.3.2 Prise en compte des couples et des chemins peu fréquents

L'utilisation de chemins syntaxiques permet de repérer de manière efficace les couples d'EN en une relation donnée. En effet, les couples d'EN reliées par une relation syntaxique entretiennent également une relation sémantique avec une forte probabilité, ce qui n'est par contre pas toujours le cas des couples d'EN repérés dans un simple n-gramme. Si bien que dans les travaux de Hasegawa, les couples de fréquence peu élevée n'ont pas été exploités. La bonne précision que nous avons eue, tout en traitant des couples de fréquence très peu élevée, montre bien la fiabilité des relations entre deux EN reliées par un chemin syntaxique.

5.3.3 Performance des chemins syntaxiques pour le repérage des relations

Comme nous l'avons déjà indiqué, nous avons fixé la longueur maximale des chemins de relations à cinq nœuds intermédiaires. Cette longueur est semblable à celle utilisée dans les travaux antérieurs pour la fenêtre d'analyse. Cependant, les relations pouvant être extraites avec ce seuil diffèrent largement avec notre méthode. Par exemple, le chemin de longueur 1 tel que : $\rightarrow_{\text{CPL-V(par)}}$ dirigée $\rightarrow_{\text{MOD-N}}$ relie aussi bien les deux EN proches comme Organisation de libération de la Palestine et Mahmoud Abbas dans : *Organisation de libération de la Palestine* ***dirigée par*** *Mahmoud Abbas*, que les deux EN éloignées comme Ligue pour la démocratie nationale et Suu Kyi dans la phrase : *Avant cela, ont indiqué ces sources, Jim Webb avait rencontré, également à Naypyidaw, des représentants de la Ligue pour la démocratie nationale (LND), principale formation d'opposition en Birmanie,* ***dirigée par*** *Mme Suu Kyi.*

Nous n'avons pas besoin non plus de nous préoccuper de l'ordre d'apparition des éléments dans leur réalisation linéaire. Les chemins, les contextes, reliant deux EN n'apparaissent pas forcément entre les deux EN, mais le positionnement des différents éléments concernés dans une réalisation linéaire est complètement transparent et ne nécessite aucunement des traitements particuliers pour chaque cas. Le chemin $\leftarrow_{\text{APPOS}}$ député $\rightarrow_{\text{APPOS}}$ relie les deux EN apparaissant toutes les deux dans son contexte droit comme dans : *Le* ***député*** *PS Arnaud Montebourg a affirmé mardi que …* . De plus, le même chemin relie également deux EN apparaissant dans un ordre inverse comme : *Le* ***député*** *Patrick Braouezec (PCF)*, permettant ainsi de classer ces deux couples dans le même groupe sans aucun traitement particulier.

6. Utilisation des résultats d'extraction pour l'enrichissement d'une ontologie

6.1 Un référentiel ontologique d'entités nommées

Dans le cadre d'expériences d'enrichissement sémantique de dépêches de l'Agence France Presse, un référentiel de métadonnées pertinentes pour cet enrichissement est en cours de développement[3]. Ce référentiel est destiné à collecter les informations recueillies à partir des dépêches, à les organiser et les maintenir dans une structure cohérente afin de rendre possible et aisée leur exploitation, à des fins de catégorisation, de recherche documentaire ou de filtrage thématique des dépêches par exemple. Les entités nommées constituent les données principales de ce référentiel. Le modèle de représentation de ces connaissances est une ontologie conçue notamment pour modéliser les classes conceptuelles correspondant à une typologie classique d'entités nommées (Personnes, Lieux, Organisations). Le langage ontologique choisi est OWL-DL. Outre les classes d'entités, d'autres concepts importants pour la modélisation sont présents dans cette ontologie, comme par exemple les catégories thématiques utilisées par l'agence.

La population de ce référentiel, pour ce qui concerne les classes principales (les EN), se fait progressivement par extraction d'entités sur les corpus de l'AFP ; ces entités, une fois validées manuellement, instancient une des classes d'entités suivant leur type. Cet ensemble croissant d'instances d'entités présente par ailleurs un certain nombre de connaissances représentées sous forme d'attributs (variantes du nom, nom canonique, page Web correspondante…) ou de relations avec d'autres entités (la nationalité d'une instance de personne est une relation avec une instance de la classe *Pays*, par exemple).

Cependant, des informations plus complexes et soumises au changement dans le temps peuvent conduire à une maintenance difficile d'une telle ontologie. Il s'agit notamment de relations entre entités comme l'appartenance d'une personne à une organisation, la direction d'une entreprise par une personne, etc. Le langage OWL permet l'instanciation d'un *ObjectProperty* afin de rendre compte d'une connaissance de ce type, ce qui nécessiterait la création d'un nouvel *ObjectProperty* lors de chaque nouvelle relation à intégrer dans le référentiel. Bien qu'envisageable, ce procédé peut se révéler lourd en termes de gestion et de cohérence de l'ontologie. Par ailleurs, le choix peut être fait, dans la conception et la maintenance d'une ontologie, de fixer au préalable l'étendue des classes et relations instanciables, et de la limiter autant que la maintenance peut l'exiger. Ainsi, l'administration de l'ontologie pourra rendre impossible la création d'un nombre important d'*ObjectProperty*, chacun correspondant à une relation particulière entre deux

[3] Le cadre applicatif de ces travaux est décrit dans (Stern & Sagot 2010)

entités. Cette solution a également l'inconvénient de représenter de façon fixe une connaissance de ce type : la direction d'une entreprise entre deux dates sera par exemple uniquement identifiable par le label de l'*ObjectProperty* créé *ad hoc*, comme *isDirectorOfFrom1999to2005*.

La représentation de ce type de connaissances sur les entités et leurs relations peut se faire sans l'utilisation d'*ObjectProperty* et ainsi éviter les écueils de l'explosion du nombre d'éléments ontologiques à gérer d'une part, et du peu de complexité des informations permises d'autre part. Le choix fait pour cette ontologie est en effet de réifier les relations entre entités en les faisant correspondre à des classes conceptuelles. Cela permet, d'une part, de définir à l'avance le type de relations auxquelles l'ontologie doit s'intéresser concernant les entités qui la peuplent en intégrant les relations dans une hiérarchie conceptuelle et, d'autre part, de bénéficier de la richesse de représentation propre aux instances de classes (attributs et relations).

6.2 Relations entre entités : prédicats de fonctions

Ainsi, l'ontologie considérée présente une classe dédiée à la représentation d'un type de relation entre entités : les fonctions, telles que la direction d'une entreprise ou la présidence d'une institution par une personne[4]. Il s'agit de relations typiquement instanciables entre entités de type *Person* et *Organization*. Cette classe *Function* est destinée à être instanciée par des situations de fonctions particulières, telles que *FrancePresidency* pour « présidence de la France ». Seuls deux *ObjectProperties* sont définis de façon générale comme pouvant intervenir dans les instances de cette classe : *hasFunctionFiller* et *isFunctionFilledFor*. Dans les deux cas le *domain* est une instance de la classe *Function*. La première a pour *range* une

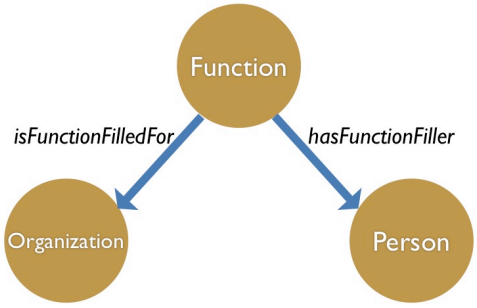

FIG. 4 – *Modélisation de la relation de fonction*

[4] L'ontologie se limite à ce jour à ce type de relations sans restriction pour de futurs développements.

instance de la classe *Person* – celle qui remplit la fonction, et la seconde a pour *range* l'instance de la classe *Organization* concernée par cette fonction. La figure 4 montre le schéma illustrant cette modélisation.

C'est donc le nombre d'instances de fonctions qui augmente au fur et à mesure que de nouvelles connaissances sont intégrées, et non le nombre de relations différentes à travers des *ObjectProperties*. La relation existant entre deux entités est donc représentée sous la forme d'un *triplet prédicatif* et est par ailleurs fortement caractérisée conceptuellement grâce à son appartenance à une classe de l'ontologie ; la classe *Function* possède en effet des sous-classes permettant de spécifier le type de fonction considérée (politique, sportive…) tout en contraignant les types possibles. Par ailleurs, des informations peuvent être ajoutées à l'instance de fonction à l'aide d'attributs ; par exemple, la validité temporelle ou historique d'une relation pourra être spécifiée par des attributs *hasStartDate* et *hasEndDate* ; notre exemple pourra être illustré ainsi :

FIG. 5 – *Exemple d'instanciation d'une fonction dans l'ontologie (interface : logiciel* Protégé*)*

6.3 Population des fonctions à partir des relations extraites sur corpus

Les travaux présentés précédemment fournissent un ensemble de connaissances sur les entités présentes dans les dépêches de l'AFP. Cet ensemble peut venir peupler l'ontologie décrite ici puisqu'il s'agit d'enrichir les informations relatives aux entités préalablement instanciées.

La population de l'ontologie se fait à partir de classes d'entités et de relations prédéfinies: entités nommées et fonctions remplies par des personnes dans des organisations. Les résultats d'extraction et les regroupements de chemins de relations obtenus sont donc tout d'abord examinés afin de déterminer quels sont

ceux pouvant correspondre aux classes de fonctions de l'ontologie. Il s'agit principalement des chemins IND-ORG (§5). Chaque classe de chemins, comme *directeur, capitaine, patron* (§5.1) est mise en relation avec une sous-classe de la classe *Function*, puis chaque réalisation de cette classe de chemins est intégrée à l'ontologie sous la forme d'une instance, selon la modélisation décrite plus haut.

Les entités concernées par ces instances de classes de chemins doivent elles aussi être liées à l'instance d'entité correspondante. Cela est possible par le biais d'une étape de *résolution*[5] des entités en regard du référentiel ontologique : chaque entité présente dans un chemin de relation extrait reçoit un identifiant propre au référentiel ; dans le cas d'une ambiguïté entre plusieurs entités de ce référentiel, la résolution utilise des informations contextuelles dans le texte d'origine et les connaissances sur les entités déjà présentes dans l'ontologie, afin d'établir une mesure de similarité entre l'entité détectée et les différentes instances du référentiel qui peuvent lui correspondre. L'entité candidate la plus similaire est ensuite sélectionnée pour recevoir l'instanciation de la relation de fonction extraite.

Cette population permet donc d'obtenir un enrichissement des connaissances disponibles sur les entités instanciées dans le référentiel ontologique. Ces connaissances ont un caractère ontologique intéressant puisqu'elles produisent un véritable réseau entre entités et apportent de ce fait une information riche et complexe au référentiel, qu'il est ensuite possible d'exploiter dans diverses applications nécessitant des raisonnements sur les entités présentes dans la production de dépêches.

7. Conclusion

Nous avons proposé une méthode non-supervisée d'extraction des relations et des patrons de relations entre entités nommées, réalisée pour l'enrichissement d'une ontologie. Nous avons également présenté le résultat d'une expérience d'évaluation de notre méthode qui montrait qu'en dépit d'un nombre bien supérieur d'instances extraites de différentes relations, sa performance était tout aussi bonne que celle des travaux antérieurs. Une des pistes d'amélioration résiderait dans la possibilité de modification de notre formule de calcul de similarité des chemins pour la classification afin de réduire le nombre de classes formées. Les résultats obtenus par notre méthode sont en cours d'intégration dans une ontologie modélisant des connaissances relatives à des entités nommées.

[5] Le module de résolution est un élément de la chaîne décrite dans (Stern & Sagot 2010-1 et 2010-2).

Références

Agichtein, E. & Gravano, L. (2000). Snowball : Extracting relations from large plain-text collections. In Proceedings of the 5th ACM International Conference on Digital Libraries, p. 85-94.

Banko M., Cafarella M. J., Soderland S., Broadhead M. & Etzioni O. (2007). Open information extraction from the web. In IJCAI'07, p. 2670–2676.

Bollegala D., Matsuo Y. & Ishizuka M. (2010). Relational duality : Unsupervised extraction of semantic relations between entities on the web. In Proc. of the 19th International Conference on World Wide Web (WWW 2010), p. 151–160.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98, p. 172-183.

Bunescu, R. & Mooney, R. (2005). A shortest path dependency kernel for relation extraction. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, p. 724-731, Vancouver, British Columbia, Canada : Association for Computational Linguistics.

Chen, J., Ji, D.-H., Tan, C. L. & Niu, Z.-Y. (2005). Automatic relation extraction with model order selection and discriminative label identification. In IJCNLP, p. 390-401.

Hasegawa, T., Sekine, S. & Grishman, R. (2004). Discovering relations among named entities from large corpora. In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume, p. 415-422, Barcelona, Spain.

He, T., Zhao, J. & Li, J. (2006). Discovering relations among named entities by detecting community structure. In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, p. 42-48.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International Conference on Computational Linguistics, p. 539-545.

Iftene A. & Balahur-Dobrescu A. (2008). Named entity relation mining using wikipedia. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco : European Language Resources Association (ELRA).

Lin D. & Pantel P. (2001) Discovery of Inference Rules for Question Answering. In Natural Language Engineering, 7(4), p. 343-360.

Matsuo Y., Mori J., Hamasaki M., Ishida K., Nishimura T., Takeda H., Hashida K. & Ishizuka M. (2006). Polyphonet : An advanced social network extraction system from the web. In Proc. of the 15th International Conference on World Wide Web (WWW 2006).

Nakamura-Delloye, Y. & Villemonte de la Clergerie, E. (2010). Exploitation de résultats d'analyse syntaxique en vue d'acquisition de relations entre entités nommées. In Actes de TALN 2010 (Traitement automatique des langues naturelles), Montréal, Canada.

Sagot B. & Boullier P. (2008). SXPipe 2 : architecture pour le traitement présyntaxique de

corpus bruts. In Traitement Automatique des Langues (T.A.L.), 49(2), p. 155–188.

Stern R. & Sagot B. (2010-1). Resources for named entity recognition and resolution in news wires. In Proceedings of LREC 2010 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management, La Valette, Malte.

Stern R. et Sagot B. (2010-2). Détection et résolution d'entités nommées dans des dépêches d'agence. In Actes de TALN 2010, Montréal, Canada.

Villemonte de la Clergerie, E. (2010). Convertir des dérivations TAG en dépendances. In Actes de TALN 2010, Montréal, Canada.

Villemonte de la Clergerie, E., Sagot, B., Nicolas, L. & Guénot, M.-L. (2009). FRMG : évolutions d'un analyseur syntaxique tag du français. In Journée ATALA "Quels analyseurs syntaxiques pour le français ?".

Zelenko, D., Aone, C. & Richardella, A. (2002). Kernel methods for relation extraction. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), p. 71-78.

Zhang, M., Su, J., Wang, D., Zhou, G. & Tan, C. L. (2005). Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In IJCNLP, p. 378-389.

Annexe

Les travaux décrits dans cet article ont débuté avec le projet SCRIBO (Semi-automatic and Collaborative Retrieval of Information Based on Ontologies), labellisé par le pôle de compétitivité System@tic et financé par la DGE, et se sont poursuivis dans le cadre du projet EDyLex (Enrichissement Dynamique de ressources Lexicales multilingues en contexte multimodal), financé par l'ANR (ANR-09-CORD-008).
Site internet du projet : http://sites.google.com/site/projetedylex/.

Summary

We propose in this paper an unsupervised method for relation and pattern extraction. Our work is carried out under an ontology building and enrichment project. The proposed method is characterized by using parsed corpora, especially by leveraging syntactic paths that connect two named entities in dependency trees. Information on the syntactic relations between constituents is used to improve the similarity calculation for the clustering. We also describe how to integrate the obtained results in our ontology.